Biometrika (2016), **xx**, x, *pp*. 1–17

Printed in Great Britain

In-sample forecasting with local linear survival densities

BY M. HIABU
Cass Business School, City University London, 106 Bunhill Row, London EC1Y8TZ, U.K.
munir.hiabu.1@cass.city.ac.uk

E. MAMMEN
Institute for Applied Mathematics, Heidelberg University, INF 205, 69120 Heidelberg, Germany
mammen@math.uni-heidelberg.de

M. D. MARTÍNEZ-MIRANDA
Cass Business School, City University London, 106 Bunhill Row, London EC1Y8TZ, U.K.
lola.martinez-miranda@city.ac.uk

AND J. P. NIELSEN
Cass Business School, City University London, 106 Bunhill Row, London EC1Y8TZ, U.K.
jens.nielsen.1@city.ac.uk

SUMMARY

In this paper, in-sample forecasting is defined as forecasting a structured density to sets where it is unobserved. The structured density consists of one-dimensional in-sample components that identify the density on such sets. We focus on the multiplicative density structure, which has recently been seen as the underlying structure of non-life insurance forecasts. In non-life insurance the in-sample area is defined as one triangle and the forecasting area as the triangle that added to the first triangle produces a square. Recent approaches estimate two one-dimensional components by projecting an unstructured two-dimensional density estimator onto the space of multiplicatively separable functions. We show that time-reversal reduces the problem to two one-dimensional problems, where the one-dimensional data are left-truncated and a one-dimensional survival density estimator is needed. This paper then uses the local linear density smoother with weighted cross-validated and do-validated bandwidth selectors. Full asymptotic theory is provided, with and without time reversal. Finite sample studies and an application to non-life insurance are included.

Some key words: Aalen's multiplicative model; Cross-validation; Do-validation; Density estimation; Local linear kernel estimation; Survival data.

1. INTRODUCTION

This paper develops a dimension-reduction procedure in order to forecast an age-cohort structure. Our motivating example is taken from non-life insurance where the estimation of outstanding liabilities involves an age-cohort model. In non-life insurance such a structure is called chain-ladder: cohorts are based on the year of underwriting the insurance policy and age is the devel-

opment of claims. Age-cohort and chain-ladder models have often been formulated as discrete models aggregating observations in months, quarters or years. Martínez-Miranda et al. (2013) identified the chain-ladder method as a structured histogram in the vocabulary of non-parametric smoothing, and suggested replacing the structured histogram smoothers by continuous kernel smoothers, which are more efficient.

We assume that our data follow a joint distribution with independent components, one for cohort and one for age, but are truncated if cohort plus age is greater than the calendar time of data collection. Future observations remain unobserved, and the forecasting exercise is to predict them. Visualised, the historical data belong to a triangle and the forecasting exercise is to predict the densities on the triangle that added to the first completes a square. We call this forecasting structure in-sample forecasting, because information on the two relevant densities of the multiplicative structure is indeed in the sample. The independence assumption for the unfiltered data will be discussed in the next section. Our model is thus that we have independent and identically distributed truncated observations sampled from the two-dimensional random variable, (X, Y), with values on the triangle $\mathcal{I} = \{(x, y) : \ x + y \leq T, \ x, y \geq 0\}$, $T \in \mathbb{R}_+$. These observations are truncated from the complete set with support on the square $[0, T]^2$. We wish to make in-sample forecasts of the density with support on the second triangle, $\mathcal{J} = [0, T]^2 \setminus \mathcal{I}$, which completes the square. Furthermore, for unfiltered (X, Y), the joint density, f, has support on the whole square, $[0, T]^2$ and is multiplicative, i.e., $f(x, y) = f_1(x)f_2(y)$. Given this multiplicative structure, the truncated observations provide in-sample information about the density in the forecasting triangle. Estimating only the survival functions or cumulative hazards is not enough when integrating the forecasts considered in this paper, since $\mathcal{J}$ is non-rectangular.

We estimate the two multiplicative components without first having to estimate the two-dimensional density. This is possible due to the reinterpretation of the forecasting aim as two distinct one-dimensional right-truncated density estimation problems, which can be solved in a counting process framework. It is well-known that intractable right-truncation can be replaced by more tractable left-truncation by reversing the time scale; see for example Ware & DeMets (1976) and Lagakos et al. (1988). The time-reversal approach requires estimates of the survival densities, for which we use the local linear survival kernel density estimator of Nielsen et al. (2009) with cross-validated or do-validated bandwidths; see Mammen et al. (2011), Gámiz et al. (2013) and Gámiz et al. (2016). We introduce full asymptotic theory of the corresponding bandwidth selectors with and without weighting, and with and without time reversal. Reducing the forecasting to a one-dimensional problem enables us to introduce a new measure of forecasting accuracy that is equivalent to an importance-weighted loss function. The bandwidths chosen by this new measure focus on the areas of the one-dimensional functions that are most important for the forecast. When estimating outstanding liabilities, least information is available for the most recent years but they are the most important ones to estimate accurately. The new approach leads to larger bandwidths than classical goodness-of-fit loss measures. This better reflects the nature of the underlying problem, and improves forecasting accuracy.

2. IN-SAMPLE FORECASTING AND RELATED WORK

While we use counting process theory in this paper to reduce the number of dimensions, the problem can also be formulated via independent stochastic variables X and Y and their density on a triangular support; see Martínez-Miranda et al. (2013), Mammen et al. (2015) and Lee et al. (2015), where in the two latter papers the triangular support is one special case. The density components of X and Y have direct analogues in survival analysis. The density f_1 of X measures exposure, i.e., the number of individuals at risk, while the density f_2 of Y corresponds

to duration. While classical counting process theory in survival analysis operates with observed exposure, in-sample forecasting estimates f_1 and does not need observed exposure. This has the advantage of requiring fewer data. Simple model assumptions are often preferable when forecasting, so in-sample forecasting might be preferable even in situations where more data, including exposure, are available.

For example when reserves for outstanding liabilities are to be estimated in insurance companies, there is usually no follow-up data of individuals in the portfolio available and reported claims, categorised in different businesses and other baseline characteristics, are the only records. The reason that insurers do not use classical biostatistical exposure data, i.e., they do not follow every underwritten policy, might be because of the bad quality and complexity of such exposure data, with many potential causes of failure which heavily affect the actual cost of a claim.

When claim numbers are considered, then X is the underwriting date of the policy, and Y is the time between underwriting date and the report of a claim, the reporting delay. Truncation occurs when $X + Y$ is smaller than the date of data collection. The mass of the unobserved, future triangle, $\mathcal{J}$, then corresponds to the proportion of claims underwritten in the past which are not yet reported. The assumption of a multiplicative density means that the reporting delay does not depend on the underwriting date. Thus, calendar time effects like court rulings, emergence of latent claims, or changes in operational time cannot be accommodated in the model per se. Nevertheless we restrict our discussion to the multiplicative model for several reasons. It has its justification as baseline for generalisations in many directions. It also approximates the data structure well enough in many applications. We will return to this point when discussing our data example. The relevance of the multiplicative model also lies in the fact that it helps the understanding of related discrete versions that are used in all non-life insurance companies; see England & Verrall (2002).

The underlying model before filtering is the same in forward and backward time, namely that the underlying sampled random variables, X and Y, are independent with joint multiplicative density $f(x, y) = f_1(x)f_2(y)$. This multiplicative structure based on partially observed independent random variables is well known in biostatistical theory and can be checked via independence tests of Tsai (1990), Mandel & Betensky (2007) and Addona et al. (2012). Brookmeyer & Gail (1987) aimed at understanding the estimation of outstanding numbers of onset AIDS cases from a given population. They considered prevalent cohorts, where time of origin is not known, and discussed the resulting biases from just using the prevalent time available instead of infection time of each observed individual. Wang (1989) works with prevalent cohort data, but where time of origin is known, and points out that this sampling boils down to a random truncation model. Both papers work in forward-moving time but could have used the filtered non-parametric density approach of this paper, see §6, had it existed.

In the in-sample forecasting application two sampling details are different, leading us to reverse the time and using the non-parametric density approach in reversed time. One is that less is known than in Wang (1989), because exposure, i.e., the number of people at risk, is unobserved. Another is that more is known than in Wang (1989), because all failures are observed, without exception. In reversed time, the future numbers of failures, the past number of failures in regular time, is exactly the exposure needed for estimation. Therefore, the extra information that all failures are observed up to a point can alleviate the challenge of unobserved exposure, and the technique doing this is to reverse the direction of time. The most favourable approach depends on the application and situation. For approaches that study reserving for outstanding liabilities and that work with exposure in forward-moving time see Arjas (1989), Norberg (1993) and Antonio & Plat (2014).

3. MODEL

Consider the probability space $\{\mathcal{S}, \mathcal{B}(\mathcal{S}), P\}$, where $\mathcal{S}$ is the square $\{(x,y) : 0 \leq x, y \leq T\}$, where $\mathcal{B}(\cdot)$ denotes the Borel σ-field. We are interested in estimating the density, $f = \mathrm{d}P/\mathrm{d}\lambda$, where λ is the two-dimensional Lebesgue measure. We will assume that $f(x,y) = f_1(x)f_2(y)$, and that observations are only sampled on a subset of the full support of this density, f. The truncated density is assumed to be supported on the triangle, $\mathcal{I}$. In this case, we consider observations of the independent and identically distributed pairs, $\{(X_1, Y_1), \ldots, (X_n, Y_n)\}$, with $X_i \leq T - Y_i$, or equivalently $Y_i \leq T - X_i$, where T is the calendar time at which the data are collected. Since the observations are truncated, and hence $\mathcal{I}$-valued, the pair (X_1, Y_1) is not distributed according to P, but has density $f(x,y)/P(\mathcal{I})$. Both observation schemes can be understood as random right-truncation targeting only X or Y, respectively, and so both can be formulated in the following counting-process framework. We define two counting processes, one indicating the occurrence of X, and the other indicating the occurrence of Y. By reversing the times of the counting processes, right-truncation becomes left-truncation (Lagakos et al., 1988).

We define the two time reversed counting processes as

$$N_1^i(t) = I\left(T - X_i \leq t\right), \quad N_2^i(t) = I\left(T - Y_i \leq t\right) \quad (i = 1, \ldots, n),$$

with respect to the filtrations

$$\mathcal{F}_{1,t}^i = \sigma\left(\left\{T - X_i \leq s : \ s \leq t\right\} \cup \left\{Y_i \leq s : \ s \leq t\right\} \cup \mathcal{N}\right),$$
$$\mathcal{F}_{2,t}^i = \sigma\left(\left\{T - Y_i \leq s : \ s \leq t\right\} \cup \left\{X_i \leq s : \ s \leq t\right\} \cup \mathcal{N}\right),$$

satisfying the usual conditions (Andersen et al., 1993, p. 60), and where $\mathcal{N} = \{A : \ A \subseteq B, \ B \in \mathcal{B}(\mathcal{S}), P(B) = 0\}$. Adding the null set, $\mathcal{N}$, to the filtration guarantees its completeness. This is a technically useful construction, but it has been argued that it is not necessary; see Jacod (1979) and Jacod & Shiryaev (1987). We keep the assumption because we use results that rely on it.

Both counting processes operate on a reversed timescale, so all the usual estimators derived from these counting processes will be based on $T - X$ and $T - Y$, rather than on X and Y. To minimize any potential confusion, we will mark all functions corresponding to $T - X$ or $T - Y$ with a superscript, R. The desired estimators will then be linear transformations of the time-reversed versions.

The advantage of this time reversal can be seen by identifying the random intensity of N_l^i, λ_l^i, which is well-defined since X and Y have bounded densities. It holds, almost surely, that for all $t \in [0,T]$ we have $\lambda_l^i(t) = \lim_{h \downarrow 0} h^{-1} E\left[N_l^i\left\{(t+h)-\right\} - N_l^i(t-) \middle| \mathcal{F}_{t-}\right]$ $(l = 1, 2)$, see Aalen (1978). At this point we used that our assumptions will imply that λ_l^i is piecewise continuous. Straightforward computations lead to intensities satisfying Aalen's (1978) multiplicative intensity model,

$$\lambda_l^i(t) = \alpha_l(t) Z_l^i(t),$$

with predictable processes $Z_1^i(t) = I\big(Y_i < t \leq T - X_i\big)$, $Z_2^i(t) = I\big(X_i < t \leq T - Y_i\big)$, cumulative distribution functions $F_l = \int_0^{\cdot} f_l(x)\mathrm{d}x$ $(l = 1, 2)$, and hazard ratios

$$\alpha_1(t) = \lim_{h \downarrow 0} h^{-1} \mathrm{pr}\left\{T - X \in [t, t+h) \mid \ T - X \geq t\right\} = \frac{f_1(T-t)}{F_1(T-t)} = \frac{f_1^R(t)}{S_1^R(t)},$$
$$\alpha_2(t) = \lim_{h \downarrow 0} h^{-1} \mathrm{pr}\left\{T - Y \in [t, t+h) \mid \ T - Y \geq t\right\} = \frac{f_2(T-t)}{F_2(T-t)} = \frac{f_2^R(t)}{S_2^R(t)}.$$

As the hazard function, α_1, does not depend on f_2, and the hazard function, α_2, does not depend on f_1, we can estimate f_1 and f_2 as one-dimensional densities.

4. LOCAL LINEAR DENSITY ESTIMATOR IN REVERSED TIME

Due to the symmetry between $T - X$ and $T - Y$, all of the following results hold for both f_1 and f_2. For clarity, therefore, we suppress the subscript l, which indicates the coordinate. Furthermore, we will denote the exposure or risk process by $Z(t) = \sum_{i=1}^{n} Z^i(t)$.

Following Nielsen et al. (2009), our proposed estimator of the density function, f^R, will involve a pilot estimator of the survival function, $S^R(t)$. Here, for simplicity, we choose the Kaplan–Meier product-limit estimator, $\widehat{S}^R(t) = \prod_{s \le t} \left\{1 - \Delta\widehat{\mathrm{A}}(s)\right\}$, where $\widehat{A}(t) = \sum_{i=1}^{n} \int_0^t \{Z(s)\}^{-1} \mathrm{d}N^i(s)$ is the Aalen estimator of the integrated hazard function, $A(t) = \int_0^t \alpha(s)\mathrm{d}s$. We define the local linear estimator $\widehat{f}^R_{h,K}(t)$ of $f^R(t)$ as the minimizer $\widehat{\theta}_0$ in the equation

$$\begin{pmatrix} \widehat{\theta}_0 \\ \widehat{\theta}_1 \end{pmatrix} = \arg \min_{\theta_0, \theta_1 \in \mathbb{R}} \sum_{i=1}^{n} \left[\int K_h(t-s) \left\{\theta_0 + \theta_1(t-s)\right\}^2 Z^i(s) W(s) \mathrm{d}s \right.$$
$$\left. - 2 \int K_h(t-s) \left\{\theta_0 + \theta_1(t-s)\right\} \widehat{S}^R(s) Z^i(s) W(s) \mathrm{d}N^i(s) \right]. \quad (1)$$

Here and below, an integral $\int$ with no limits denotes integration over the whole support, i.e., $\int_0^T$. In addition, for kernel K and bandwidth h, $K_h(t) = h^{-1}K(t/h)$. The definition of the local linear estimator as the minimizer of (1) can be motivated by the fact that the sum on the right hand side of (1) equals the limit of

$$\sum_{i=1}^{n} \int \left[\left\{ \frac{1}{\varepsilon} \int_s^{s+\varepsilon} \widehat{S}^R(u) \mathrm{d}N^i(u) - \theta_0 - \theta_1(t-s) \right\}^2 - \xi(\varepsilon) \right] K_h(t-s) Z^i(s) W(s) \, \mathrm{d}s,$$

for ε converging to zero. Here, $\xi(\varepsilon) = \{\varepsilon^{-1} \int_s^{s+\varepsilon} \widehat{S}^R(u) \mathrm{d}N^i(u)\}^2$ is a vertical shift subtracted to make the expression well-defined. Because $\xi(\varepsilon)$ does not depend on (θ_0, θ_1), $\widehat{\theta}_0$ is defined by a local weighted least squares criterion. The function W is an arbitrary predictable weight function on which the pointwise first order asymptotics will not depend. There exist two popular weightings: the natural unit weighting, $W(s) = 1$, and the Ramlau–Hansen weighting, $W(s) = \{n/Z(s)\} I\{Z(s) > 0\}$. The latter becomes the classical kernel density estimator in the simple unfiltered case. However, in the framework of filtered observations the natural unit weighting, $W(s) = 1$, tends to be more robust (Nielsen et al., 2009), so we use it. For this, the solution of (1) (Nielsen et al., 2009; Gámiz et al., 2013) is

$$\widehat{f}^R_{h,K}(t) = n^{-1} \sum_{i=1}^{n} \int \overline{K}_{t,h}(t-s) \widehat{S}^R(s) \mathrm{d}N^i(s), \quad (2)$$

where

$$\overline{K}_{t,h}(t-s) = \frac{a_2(t) - a_1(t)(t-s)}{a_0(t)a_2(t) - \{a_1(t)\}^2} K_h(t-s),$$
$$a_j(t) = n^{-1} \int K_h(t-s)(t-s)^j Z(s) \mathrm{d}s \quad (j = 0, 1, 2).$$

If K is a second-order kernel, then $n^{-1}\int \overline{K}_{t,h}(t-s)Z(s)\mathrm{d}s = 1$, $n^{-1}\int \overline{K}_{t,h}(t-s)(t-s)Z(s)\mathrm{d}s = 0$, $n^{-1}\int \overline{K}_{t,h}(t-s)(t-s)^2Z(s)\mathrm{d}s > 0$, so that $\overline{K}_{t,h}$ can be interpreted as a second-order kernel with respect to the measure, μ, where $\mathrm{d}\mu(s) = n^{-1}Z(s)\mathrm{d}s$. This is essential in understanding the pointwise asymptotics of the local linear estimator which, as we will see, coincides with the kernel estimator $\sum_{i=1}^n \int K_h(t-s)\widehat{S}^R(s)\{Z_1(s)\}^{-1}\mathrm{d}N^i(s)$. The stochastic character of the local linear kernel allows for the automatic adjustment near boundary regions; see also Fan & Gijbels (1996), Nielsen (1998) and Nielsen et al. (2009).

We introduce the following notation. For every kernel, K, let

$$\mu_j(K) = \int s^j K(s)\mathrm{d}s, \quad R(K) = \int K^2(s)\mathrm{d}s, \quad \overline{K}^*(u) = \frac{\mu_2(K) - \mu_1(K)u}{\mu_2(K) - \{\mu_1(K)\}^2}K(u).$$

We make the following assumptions.

Condition 1. The bandwidth $h = h(n)$ satisfies $h \to 0$ and $n^{1/4}h \to \infty$ for $n \to \infty$.

Condition 2. The densities f_1 and f_2 are strictly positive and twice continuously differentiable.

Condition 3. The kernel K is symmetric, has bounded support and has finite second moment.

Conditions 2 and 3 are standard in smoothing theory. In contrast to the unfiltered case, Condition 1 assumes more than just the bandwidth h converging to zero. This is required, otherwise the estimation error of the survival function would determine the first-order asymptotic properties of the bias, since $n^{-1/2}/h^2 \to 0$ would not hold. We write $\widehat{f}_{h,K}(t) = \widehat{f}^R_{h,K}(T-t)$, for the local linear estimator of the density f. The key in obtaining the pointwise limit distribution of $\widehat{f}_{h,K}(t) - f(t)$ is to split the estimation error into a sum of a stable part and a martingale part,

$$B^R(t) = f^{R,*}_{h,K}(t) - f^R(t), \quad V^R(t) = \widehat{f}^R_{h,K}(t) - f^{R,*}_{h,K}(t),$$

where $f^{R,*}_{h,K}(t) = n^{-1}\sum_{i=1}^n \int \overline{K}_{t,h}(t-s)Z^i(s)\widehat{S}^R(s)\alpha(s)\,\mathrm{d}s$. The estimation error can then be described as

$$\widehat{f}_{h,K}(t) - f(t) = B^R(T-t) + V^R(T-t) = B(t) + V(t).$$

PROPOSITION 1. *Under Conditions 1–3, for* $t \in (0,T)$,

$$(nh)^{1/2}\left\{\widehat{f}_{l,h,K}(t) - f_l(t) - B_l(t)\right\} \to N\left\{0, \sigma_l^2(t)\right\} \quad (l = 1,2), \quad n \to \infty,$$

in distribution, where $B_l(t) = \frac{1}{2}\mu_2(\overline{K}^*)f_l''(t)h^2 + o(h^2)$, $\sigma_l^2(t) = \lim_{n\to\infty} nh$ $\langle V_l\rangle_t = R(\overline{K}^*)f_l(t)F_l(t)\gamma_l(t)^{-1}$, $\gamma_l(t) = \mathrm{pr}(Z_l^1(t) = 1)$.

Proposition 1 is proved in the Supplementary Material.

5. BANDWIDTH SELECTION IN REVERSED TIME

5·1. *Cross-validation and do-validation*

For a kernel estimator, the bandwidth is a positive scalar parameter controlling the smoothing degree. Data-driven cross-validation in density estimation goes back to Rudemo (1982) and Bowman (1984). Nowadays, a slightly modified version (Hall, 1983) is used intended to minimize the integrated squared error. By adding a general weighting, w, and the exposure, Z, which acknowledges the filtered observations, the aim is to find the minimizer of the integrated

squared error $\Delta_K(h) = \int \left\{\widehat{f}^R_{h,K}(t) - f^R(t)\right\}^2 Z(t)w(t)\,\mathrm{d}t$, which has the same minimizer as $\int\{\widehat{f}^R_{h,K}(t)\}^2 Z(t)w(t)\,\mathrm{d}t - 2\int \widehat{f}^R_{h,K}(t)f^R(t)Z(t)w(t)\,\mathrm{d}t$. Only the second integral of this term needs to be estimated. For the survival density estimator defined in §4, Nielsen et al. (2009) propose choosing the bandwidth estimator, $\widehat{h}^K_{\mathrm{CV}}$, as the minimizer of

$$\widehat{Q}_{K,w}(h) = \int \left\{\widehat{f}^R_{h,K}(t)\right\}^2 Z(t)w(t)\,\mathrm{d}t - 2\sum_{i=1}^{n}\int \widehat{f}^{R,[i]}_{h,K}(t)\widehat{S}^R(t)w(t)\,\mathrm{d}N^i(t), \qquad (3)$$

where $\widehat{f}^{R,[i]}_{h,K}(t) = n^{-1}\sum_{j\neq i}\int \overline{K}_{t,h}(t-s)\widehat{S}^R(s)\,\mathrm{d}N^j(s)$. This can be seen as a generalization of classical cross-validation.

Over the last 20 years, many methods have been developed to improve cross-validation; see Heidenreich et al. (2013). One of the strongest bandwidth selectors of this review is so-called one-sided cross-validation (Hart & Yi, 1998; Martínez-Miranda et al., 2009). Under mild regularity conditions, there exists an asymptotically optimal bandwidth. However, this bandwidth is infeasible in practice, since it depends on the unknown quantities f, γ; see (8) below. One-sided cross-validation uses the fact that the ratio of asymptotically optimal bandwidths of two estimators with different kernels, K and L, is a feasible factor, $\rho(K,L) = \{R(K)\mu_2^2(L)/\mu_2^2(K)R(L)\}^{1/5}$, which depends only on the two kernels; see (8) and (9). The authors replace the kernel K used for the kernel estimator in (2), by its right-sided version $L = K_R = 2K(\cdot)I(\cdot \geq 0)$ when minimizing (3) and multiply the resulting cross-validation bandwidth by the feasible factor, $\rho(K, K_R)$, to derive a bandwidth for a kernel estimator with kernel, K. Hence, one firstly performs cross-validation with a kernel K_R in order to derive a bandwidth estimate. After correction with the factor $\rho(K, K_R)$, the bandwidth is then used in conjunction with the original kernel K to derive an estimator of the density. Such a construction makes sense if cross-validation for a one-sided kernel estimator works better than cross-validating with the original kernel K. One can generalize this idea by defining indirect cross-validation as a method where a kernel, L, can be arbitrarily chosen. We denote such bandwidth estimator by $h^L_{\mathrm{ICV}} = \rho(K,L)h^L_{\mathrm{CV}}$.

Savchuk et al. (2010) propose an indirect cross-validation procedure where one chooses a linear combination of two Gaussian kernels as kernel, L. Mammen et al. (2011) introduce the do-validation method, which performs indirect cross-validation twice by using two one-sided kernels, $L_1 = K_L = 2K(\cdot)I(\cdot \leq 0)$ and $L_2 = K_R$, as indirect kernels in (3). The do-validation bandwidth is the average of the two resulting bandwidths, $h_{\mathrm{DO}} = 0.5(h^{K_L}_{\mathrm{ICV}} + h^{K_R}_{\mathrm{ICV}})$. Cross-validation for kernels K_L and K_R works better than for K because the asymmetry of the kernels K_L and K_R leads to larger optimal bandwidths. An empirical study in favour of do-validation in our survival setting has been performed in Gámiz et al. (2013). Asymptotic theory for weighted and unweighted cross-validation and do-validation, with and without time reversal, is developed in §6 in our general survival density framework. Below we discuss how the weighting, w, in (3) can be chosen when the aim is to estimate outstanding loss liabilities.

5·2. *Weighting for application in claims reserving*

In Gámiz et al. (2013), standard cross-validation is defined as the minimizer of (3) with $w(t) = 1$. Hence, standard cross-validation can be formulated as an in-sample technique, which aims to estimate the optimal bandwidth for the estimator calculated from the given sample. However, the situation in the forecasting problem motivating this paper is different, since our interest focuses on the unobserved region.

In this section, we illustrate how to choose a reasonable weighting scheme to estimate the outstanding liabilities for a non-life insurance company. The most relevant data for this relate to the most recent time-periods, for which only a small number of data are available. This is a well-known challenge for actuaries, who generally tackle it by using expert opinion and manual adjustments to the data. Bornhuetter & Ferguson (1972), Mack (2008) and Alai et al. (2010) give a flavour of the Bornhutter–Ferguson method used by actuaries. Our smoothing methodology, based on continuous data, could be used as an alternative to these less rigorous approaches, and so replace expert opinion and manual adjustments by using information from relevant neighbourhoods according to an optimal smoothing criteria.

Unfortunately, the trivial weighting, $w = 1$, implies that the recent years only have small influence on the size of the bandwidth, due to the lack of sufficient data. In contrast, we want the weighting, $w(t)$, to depend on the estimated size of the liabilities at t, in order to give greatest weight to the most recent period. Assume that T is an integer indicating for instance months or years, then for a period, $p = 1, \dots, T$, the reserve, $R(p)$, is given as $R(p) = n\int_{p-1}^{p} f_1(s)S_2(T-s)\,\mathrm{d}s / \int_{\mathcal{I}} f(x,y)\,\mathrm{d}x\mathrm{d}y$, which is proportional to $\int_{p-1}^{p} f_1(s)F_2^R(s)\,\mathrm{d}s$. Hence if this is the quantity of interest, for short periods, we propose the following weighted integrated squared error to be the optimality criteria for estimating f_1,

$$\begin{aligned}\Delta_{1,K}(h) &= n^{-1}\int \left\{f_1(s)F_2^R(s) - \widehat{f}_{1,h,K}(s)\widehat{F}_2^R(s)\right\}^2 \,\mathrm{d}s \\ &= n^{-1}\int \left\{f_1^R(s)S_2(s) - \widehat{f}_{1,h,K}^R(s)\widehat{S}_2(s)\right\}^2 \,\mathrm{d}s. \end{aligned} \tag{4}$$

The estimator $\widehat{S}_2$ converges to S_2 uniformly with rate $n^{-1/2}$ (Andersen et al., 1993, p. 261), which is faster than the non-parametric convergence rate of the density; see Proposition 1. Thus, we can substitute $S_2(s)$ by its estimator $\widehat{S}_2(s) = 1 - \widehat{S}_2^R(T-s)$, and define

$$\widetilde{\Delta}_{1,K}(h) = n^{-1}\int \left\{f_1^R(s) - \widehat{f}_{1,h,K}^R(s)\right\}^2 \left\{\widehat{S}_2(s)\right\}^2 \,\mathrm{d}s.$$

But, since f_1 and $\widehat{S}_2$ do not depend on h, minimizing $\widetilde{\Delta}_{1,K}$ in h is equivalent to minimizing

$$\begin{aligned} Q_K(h) &= \widetilde{\Delta}_{1,K}(h) - \int \left\{f_1^R(t)\widehat{S}_2(t)\right\}^2 \,\mathrm{d}t \\ &= \int \left\{\widehat{f}_{1,h,K}^R(t)\right\}\left\{\widehat{S}_2(t)\right\}^2 \,\mathrm{d}t - 2\int f_1^R(t)\widehat{f}_{1,h,K}^R(t)\left\{\widehat{S}_2(t)\right\}^2 \,\mathrm{d}t. \end{aligned}$$

Therefore, we choose the weight $w_1(t) = \widehat{S}_2(t)^2/Z_1(t)$ in (3), and the cross-validation estimator of $Q_K(h)$ becomes

$$\begin{aligned}\widehat{Q}_{K,w_1}(h) = &\int \left\{\widehat{f}_{1,h,K}^R(t)\right\}^2\left\{\widehat{S}_2(t)\right\}^2 \,\mathrm{d}t \\ &- 2\sum_{i=1}^{n}\int \widehat{f}_{1,h,K}^{R,[i]}(t)\widehat{S}_1^R(t)\left\{\widehat{S}_2(t)\right\}^2\{Z_1(t)\}^{-1}\,\mathrm{d}N^i(t). \end{aligned} \tag{5}$$

By symmetry, the weighting for f_2 can be derived in a similar fashion, with $w_2(t) = \widehat{S}_1(t)^2/Z_2(t)$.

6. ASYMPTOTICS FOR WEIGHTED COMBINATIONS OF INDIRECT CROSS-VALIDATION

In this section we formulate the asymptotic theory of the bandwidth selectors in the original time direction. This gives statisticians using cross-validation or do-validation with the local linear density estimator of Nielsen et al. (2009); as in Gámiz et al. (2013), the asymptotic theory needed to support their approach. We then provide the theory for the reversed time direction.

We first briefly describe the general model in the original time direction (Nielsen et al., 2009; Gámiz et al., 2013). When observing n individuals, let N_i be a $\{0,1\}$-valued counting process, which observes the failures of the ith individual in the time interval, $[0,T]$. We assume that N_i is adapted to a filtration, $\mathcal{F}_t$, which satisfies the usual conditions, see §3. We also observe the $\{0,1\}$-valued predictable process, Z_i, which equals unity when the ith individual is at risk. It is assumed that Aalen's multiplicative intensity model, $\lambda_i(t) = \alpha(t)Z_i(t)$, is satisfied. This formulation contains the case of a longitudinal study with left-truncation and right-censoring. In this case, we observe triplets (Y_i, X_i, δ_i) $(i = 1, \ldots, n)$ where Y_i is the time at which an individual enters the study, X_i is the time he/she leaves the study and δ_i is binary and equals 1 if death is the reason for leaving the study. Hence, $Y_i \leq X_i$, and the counting process formulation would be $N_i(t) = I(X_i \leq t)\delta_i$ and $Z_i(t) = I(Y_i \leq t < X_i)$.

The local linear survival density estimator in the original time direction is then defined as $\widehat{f}(t) = n^{-1}\sum_{i=1}^{n}\int \overline{K}_{t,h}(t-s)\widehat{S}(s)\ dN_i(s)$, where $\widehat{S}(s)$ is the Kaplan–Meier estimator of the survival function. The integrated squared error, $\Delta_K(h)$, and the cross-validation criterion, $\widehat{Q}_{K,w}(h)$, then become

$$\Delta_K(h) = n^{-1}\sum_{i=1}^{n}\int \left\{\widehat{f}(t) - f(t)\right\}^2 w(t)Z_i(t)\,\mathrm{d}t, \tag{6}$$

$$\widehat{Q}_{K,w}(h) = \sum_{i=1}^{n}\int \left\{\widehat{f}(t)\right\}^2 Z_i(t)w(t)\,\mathrm{d}t - 2\sum_{i=1}^{n}\int \widehat{f}^{[i]}(t)\widehat{S}(t)w(t)\,\mathrm{d}N_i(t), \tag{7}$$

where $\widehat{f}^{[i]}(t) = n^{-1}\sum_{j\neq i}\int \overline{K}_{t,h}(t-s)\widehat{S}(s)\,\mathrm{d}N_j(s)$.

We will derive the asymptotic properties of weighted combinations of indirect cross-validation bandwidths and in particular of the do-validation approach. In Lemma 1 of the Supplementary Material, we prove that the integrated squared error in (6) is uniformly asymptotically equivalent to $M_K(h) = (nh)^{-1}R(\overline{K}^*)\int f(t)S(t)w(t)dt + h^4\mu_2^2(\overline{K}^*)\int\{f''(t)/2\}^2\gamma(t)w(t)\mathrm{d}t$, which leads to the optimal deterministic bandwidth selector

$$h_{\mathrm{MISE}} = C_0 n^{-1/5}, \quad C_0 = \left\{\frac{R(\overline{K}^*)}{\mu_2^2(\overline{K}^*)}\frac{\int f(t)S(t)w(t)\mathrm{d}t}{\int f''(t)^2\gamma(t)w(t)\mathrm{d}t}\right\}^{1/5}, \tag{8}$$

where $\gamma(t) = n^{-1}E\{Z(t)\}$. We define h_{ISE} as the minimizer of (7) over the interval $I_n^* = [a_1^* n^{-1/5}, a_2^* n^{-1/5}]$, where the constants $a_2^* > a_1^* > 0$ are chosen such that $a_1^* < C_0 < a_2^*$.

We will study the asymptotic properties of the weighted combinations of indirect cross-validation selectors introduced in §5·1,

$$\widehat{h}_{\mathrm{ICV}} = \sum_{j=1}^{J} m_j\rho_j h_{\mathrm{CV}}^{L_j}, \qquad \rho_j = \rho(L_j) = \left\{\frac{R(K)\mu_2^2(L_j)}{\mu_2^2(K)R(L_j)}\right\}^{1/5}, \tag{9}$$

where L_j are arbitrary kernels and m_j are weights with $\sum_{j=1}^{J} m_j = 1$. For K symmetric, $J = 2, L_1 = K_L, L_2 = K_R$, and $m_1 = m_2 = 0.5$ we get the do-validation bandwidth estimator. We make the following assumptions.

Condition 4. Let $Z = \sum_{i=1}^{n} Z_i$. The expected relative exposure function, $\gamma(t) = n^{-1}E\{Z(t)\}$ is strictly positive on the support of w, is twice continuously differentiable, and $\sup_{s\in[0,T]} |Z(s)/n - \gamma(s)| = o_P\left\{(\log n)^{-1}\right\}$, $\sup_{s,t\in[0,T],|t-s|\le C_K h} |\{Z(t) - Z(s)\}/n - \{\gamma(t) - \gamma(s)\}| = o_P\left\{(nh)^{-1/2}\right\}$, where the constant C_K is defined in Condition 5.

Condition 5. The kernels, K and L_j $(j = 1, ..., J)$, are compactly supported, i.e., the support lies within $[-C_K, C_K]$ for some constant, $C_K > 0$. The kernels are continuous on $\mathbb{R}\backslash\{0\}$ and have one-sided derivatives that are Hölder continuous on $\mathbb{R}^- = \{x : x < 0\}$ and $\mathbb{R}^+ = \{x : x > 0\}$. Thus, there exist constants c and δ such that $|g(x) - g(y)| \le c|x - y|^\delta$ for $x, y < 0$ or $x, y > 0$ with g equal to K' or L_j' $(j = 1, ..., J)$. The left and right-sided derivatives differ at most on a finite set. The kernel K is symmetric.

Condition 6. It holds that $f \in C_2([0, T])$. The second derivative of f is Hölder continuous with exponent $\delta > 0$ and f is strictly positive.

Condition 7. There exists a function $\widetilde{w} \in C_1([0, T])$, with $\sup_{t\in[0,T]} |\widetilde{w}(t) - w(t)| = o_P(1)$.

Condition 5 is a weak standard condition on kernels. Condition 6 differs from standard smoothness conditions only by the mild additional assumption that the second derivative of the density function fulfils a Hölder condition. Condition 4 is also rather weak. In the special framework considered in §1–4, we have for $l = 1$ that $Z_i(t) = I(Y_i < t \le T - X_i) = I(Y_i < t) + I(X_i \le T - t) - 1$. This gives the stochastic expansions of Condition 4 by using uniform $n^{-1/2}$-convergence of the empirical distribution function. For longitudinal data sets, as described at the beginning of §6, and other examples, one can use exponential inequalities for empirical processes to check Condition 4.

THEOREM 1. *Under Conditions 4–7, the bandwidth selector $\widehat{h}_{\mathrm{ICV}}$ of the local linear survival density estimator in the original time direction satisfies*

$$n^{3/10}\left(\widehat{h}_{\mathrm{ICV}} - h_{\mathrm{MISE}}\right) \to N\left(0, \sigma_1^2\right), \quad n^{3/10}\left(\widehat{h}_{\mathrm{ICV}} - h_{\mathrm{ISE}}\right) \to N\left(0, \sigma_2^2\right), \quad n \to \infty,$$

where

$$\sigma_1^2 = S_1 \int \left\{ \sum_{j=1}^{J} m_j \frac{R(K)}{R(\bar{L}_j)} (H_{L_j} - G_{L_j})(\rho_j u) \right\}^2 \mathrm{d}u,$$

$$\sigma_2^2 = S_1 \int \left\{ \sum_{j=1}^{J} m_j \frac{R(K)}{R(\bar{L}_j)} (H_{L_j} - G_{L_j})(\rho_j u) - H_K(u) \right\}^2 \mathrm{d}u + S_2,$$

$$S_1 = \frac{2}{25} \frac{\int S^2(t) f^2(t) \widetilde{w}^2(t) \mathrm{d}t}{R^{7/5}(K)\mu_2^{6/5}(K)\left\{\int f''(t)^2\gamma(t)\widetilde{w}(t)\,\mathrm{d}t\right\}^{3/5}\left\{\int f(t)S(t)\widetilde{w}(t)\,\mathrm{d}t\right\}^{7/5}},$$

$$S_2 = \frac{4}{25} \frac{\int f''(t)^2 S(t) f(t) \widetilde{w}^2(t)\gamma(t)\mathrm{d}t - \int \left\{\int_t^T f''(u) f(u)\widetilde{w}(u)\gamma(u)\mathrm{d}u\right\}^2 \alpha(t)\gamma^{-1}(t)\mathrm{d}t}{R^{2/5}(K)\mu_2^{6/5}(K)\left\{\int f(t)S(t)\widetilde{w}(t)\mathrm{d}t\right\}^{2/5}\left\{\int f''(t)^2\gamma(t)\widetilde{w}(t)\mathrm{d}t\right\}^{8/5}},$$

Table 1. *The factor Ψ^K in* (10) *as comparison of asymptotic variances among bandwidth selection methods.*

Method	Epanechnikov	Quartic	Sextic
Do-validation	2.19	1.89	2.36
Cross-validation	7.42	5.87	6.99
Plug-in	0.72	0.83	1.18

and $G_K(u) = I(u \neq 0)\left\{\overline{K}^{**}(u) - \overline{K}^{**}(-u)\right\}$, *and*
$H_K(u) = I(u \neq 0)\int \overline{K}^{*}(v)\left\{\overline{K}^{**}(u+v) - \overline{K}^{**}(-u+v)\right\}\mathrm{d}v$, *with*

$$\overline{K}^{**}(u) = -\frac{\mu_2(K) - \mu_1(K)u}{\mu_2(K) - \{\mu_1(K)\}^2}\{K(u) + uK'(u)\} + \frac{\mu_1(K)u}{\mu_2(K) - \{\mu_1(K)\}^2}K(u).$$

Theorem 1 is proved in the Supplementary Material. The theorem states that the relative difference between the bandwidths $h_{\mathrm{CV}}, h_{\mathrm{MISE}}$ and h_{ISE} is in probability of order $n^{-1/10}$. This can be explained intuitively by the fact that a bounded interval contains $O(n^{1/5})$ non-overlapping subintervals of length h, and the kernel estimators are thus asymptotically independent if their argument differs by a magnitude of order $O(n^{-1/5})$. The rate $n^{-1/10} = (n^{-1/5})^{1/2}$ can then be explained by a central limit theorem.

The result generalises the asymptotic properties of do-validation established by Mammen et al. (2011) in the unfiltered case. If the observations, $X_1, \ldots, X_n$, are unfiltered, i.e., $Z_i(t) = I(t \leq X_i)$, then the Kaplan–Meier estimator becomes $\widehat{S}(t) = n^{-1}\sum_i Z_i(t)$, which implies that $\gamma(t) = S(t)$. Then, by choosing the weighting $w(t) = \widehat{S}(t)^{-1}$, the integrated squared error (6) and the cross-validation criterion (7) are identical to the unfiltered case and, thus, Theorem 1 is Theorem 1 in Mammen et al. (2011).

For a fixed kernel K and different choices of weighted indirect kernels (m_j, L_j), the variances, σ_2^2, only differ in the feasible factor

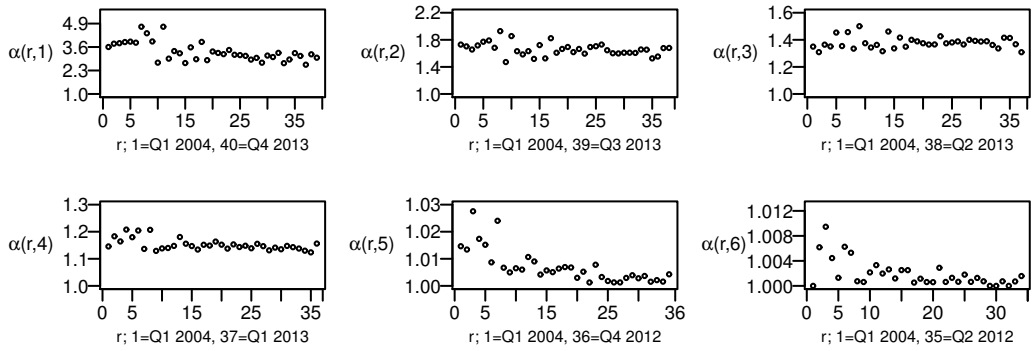
$$\Psi^K_{ICV}(m_1, \ldots, m_J, L_1, \ldots, L_J) = \int\left\{\sum_{j=1}^{J} m_j \frac{R(K)}{R(\bar{L}_j)}(H_{L_j} - G_{L_j})(\rho_j u) - H_K(u)\right\}^2 \mathrm{d}u. \tag{10}$$

The asymptotic variance of a plug-in estimation error, $(h_{\mathrm{MISE}} - h_{\mathrm{ISE}})$, is obtained by replacing the factor Ψ^K_{ICV} in σ_2^2 by $\Psi^K_{\mathrm{MISE}} = \int H_K(u)^2\,\mathrm{d}u$. Plug-in estimators are those derived by estimating the infeasible quantities of h_{MISE}, see (8), and achieve the same asymptotic limit as h_{MISE} under appropriate conditions. Their implementation in practice is not straightforward and involves pilot estimators. The values of Ψ^K can be used to compare the asymptotic performance of different methods. Table 1 shows these values for do-validation, cross-validation and the plug-in method using the Epanechnikov, quartic and sextic kernels. Once the asymptotic properties in the original time direction are derived, it is straightforward to derive a similar result in the reversed time direction.

COROLLARY 1. *Under Conditions 4–7, the bandwidth selector, $\widehat{h}_{\mathrm{ICV}}$, of the local linear survival density estimator in the reversed time direction satisfies*

$$n^{3/10}\left(\widehat{h}_{\mathrm{ICV}} - h_{\mathrm{MISE}}\right) \to N\left(0, \sigma_1^2\right), \quad n^{3/10}\left(\widehat{h}_{ICV} - h_{\mathrm{ISE}}\right) \to N\left(0, \sigma_2^2\right), \quad n \to \infty,$$

Fig. 1. Development factors of the first six quarters for individual underwriting quarters.



where

$$\sigma_1^2 = S_1 \int \left\{ \sum_{j=1}^{J} m_j \frac{R(K)}{R(\bar{L}_j)} (H_{L_j} - G_{L_j})(\rho_j u) \right\}^2 \mathrm{d}u,$$

$$\sigma_2^2 = S_1 \int \left\{ \sum_{j=1}^{J} m_j \frac{R(K)}{R(\bar{L}_j)} (H_{L_j} - G_{L_j})(\rho_j u) - H_K(u) \right\}^2 \mathrm{d}u + S_2,$$

$$S_1 = \frac{2}{25} \frac{R^{-7/5}(K) \int F^4(t) \alpha^2(T-t) \widetilde{w}^2(T-t) \mathrm{d}t}{\mu_2^{6/5}(K) \left\{ \int f''(t)^2 \gamma(T-t) \widetilde{w}(T-t)\, \mathrm{d}t \right\}^{3/5} \left\{ \int f(t) F(t) \widetilde{w}(T-t)\, \mathrm{d}t \right\}^{7/5}},$$

$$S_2 = \frac{4}{25} \left[\frac{\int f''(t)^2 F(t) f(t) \widetilde{w}^2(T-t) \gamma(T-t) \mathrm{d}t}{R^{2/5}(K) \mu_2^{6/5}(K) \left\{ \int f''(t)^2 \gamma(T-t) \widetilde{w}(T-t) \mathrm{d}t \right\}^{8/5} \left\{ \int f(t) F(t) \widetilde{w}(T-t) \mathrm{d}t \right\}^{2/5}} \right.$$
$$\left. - \frac{\int \left\{ \int_t^T f''(u) f(u) \widetilde{w}(T-u) \gamma(T-u) \mathrm{d}u \right\}^2 \alpha(t) \gamma^{-1}(t) \mathrm{d}t}{R^{2/5}(K) \mu_2^{6/5}(K) \left\{ \int f''(t)^2 \gamma(T-t) \widetilde{w}(T-t) \mathrm{d}t \right\}^{8/5} \left\{ \int f(t) F(t) \widetilde{w}(T-t) \mathrm{d}t \right\}^{2/5}} \right].$$

7. ILLUSTRATION

We now analyse a data set of reported and outstanding claims from a motor business in Cyprus. All the calculations in this and the next section have been performed with R (R Development Core Team, 2016). The data consist of $n = 58180$ claims reported between 2004 and 2013. The data are $\{(X_1, Y_1), \ldots, (X_n, Y_n)\}$, where X_i denotes the underwriting date of claim i, and Y_i the reporting delay in days. The data exist on a triangle, with $X_i + Y_i \leq 31$ December 2013. Our aim is to forecast the number of future claims from contracts underwritten in the past which have not yet been reported. It is implicitly assumed that the triangle is fully run off, such that the maximum reporting delay is ten years. This is reasonable, see Figure 2, since f_2 has a strong decay already after one year. According to the theory, we use a multiplicative structured density, $f(x, y) = f_1(x) f_2(y)$, where the components f_1 and f_2 are the underwriting date density and the development time density, respectively.

For justification of this assumption, we performed several tests, which all indicated that the assumption might be violated. We then did a more pragmatic step which is motivated from actuarial practice. We transformed the data into a triangle with dimension 3653×3653, $\mathcal{N}_{x,y} = \sum_{i=1}^{n} I\big(X_i = x,\, Y_i = y\big), \quad (x, y) \in \{1, \ldots, 3653\}^2$, and then aggregated the data into a quarterly triangle, $(\mathcal{N}^Q_{r,s})$, with dimension 40×40, which is the form usually available in a reserving department; see the Supplementary Material. For $s = 1, \ldots, 6$, we calculated the quantities $\alpha(r, s) = \sum_{l=1}^{s+1} \mathcal{N}^Q_{r,l} / \sum_{l=1}^{s} \mathcal{N}^Q_{r,l}$, known as development factors in actuarial science (Kuang et al., 2009). The values of $\alpha(r, s)$ are displayed in Figure 1. If the multiplicativity assumption is satisfied, then $\alpha(r, s)$ is approximately equal to $\{\sum_{l=1}^{s+1} f_1(x_r) f_2(y_l)\} / \{\sum_{l=1}^{s} f_1(x_r) f_2(y_l)\}$ which does not depend on r. Here, x_r lies in the rth quarter and y_l in the lth quarter. Hence, the points in each plot should lie around horizontal lines.

Only considering the first four plots, one could argue that non-constancy is caused by the stochastic nature of the observations. However, there seems to be a negative drift in the 5th and 6th plots. Non-constancy is caused in particular by the first seven underwriting quarters, which correspond to the first seven points in each plot. Re-evaluating the first four plots, one can also spot the drift there. When the values are subtracted by one, the relative drift size in the different plots seems of similar magnitude. This indicates that the data do indeed not satisfy the independence assumption, even though this is hard to see due to larger noise in the early development quarters. A pragmatic solution would be to throw away the data of the first seven underwriting quarters, as is often done by actuaries when using the chain-ladder method. We preferred to keep the whole data set because few data are available after the fourth development quarter. A better strategy might be to look for extensions of our model where the reporting delay density f_2 depends on calendar time. Additional seasonal effects are considered in Lee et al. (2015). Other calendar time effects will often involve the need of extrapolation of a time series; see also Kuang et al. (2008) for the discrete-time case. Accounting for the drift seen in the data example leads only to a slight change of the total number of forecasted claim numbers but to larger differences in the forecasted delay times.

We have calculated the local linear density estimators of the two underlying multiplicative densities, f_1 and f_2, using the Epanechnikov kernel and weighted cross-validated and do-validated bandwidth selectors. For the density f_1, cross-validation chose a bandwidth of 408 days and do-validation a bandwidth of 1,860 days, while, for f_2, the minimizer of the cross- and do-validation criteria were 15 days and 72 days, respectively. Figure 2 shows the estimated densities. The left plot indicates that there is no trend in the amount of underwritten policies. In the right plot, consistent with the policy duration of one year and our experience of other motor insurance, we find that most of the claims are reported within 1.4 years. There is a sharp increase and decrease at the beginning and at the end of the first year, respectively, and a near-uniform development in between. It seems plausible that boundary and bias correction techniques would be useful in future analyses. One could for example consider multiplicative bias correction (Nielsen et al., 2009) or asymmetric kernels (Hirukawa & Sakudo, 2014).

In this application, we encounter the usual problem with standard cross-validation, which can pick bandwidths that are much too small. Do-validation seems to have estimated a reasonable bandwidth.

The number of outstanding claims for the future quarters, obtained by integrating the multiplicative estimator over diagonals in the unobserved part, are shown in Table 2. As a benchmark, we have calculated the total reserve using the standard chain-ladder method by aggregating the data on an quarterly basis. The chain-ladder method is the most widely used reserving method in practice, and can be interpreted as a Poisson maximum likelihood estimator with multiplicative

Fig. 2. Estimated underwriting and development densities in the real data application: Cross-validation (dashed), do-validation (solid).

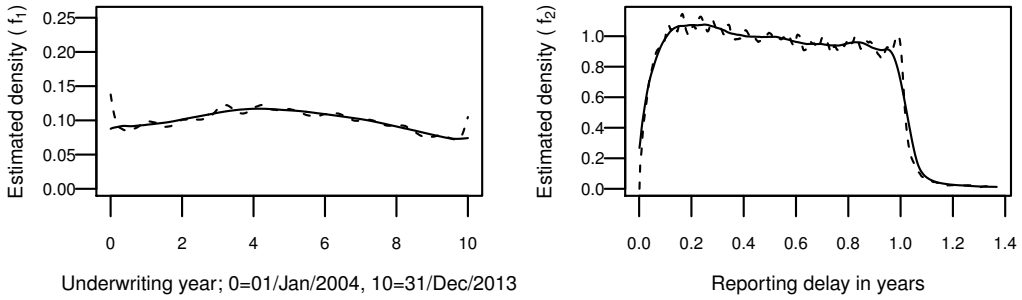


Table 2. *Number of claims forecasts in the real data application in quarters;* $1 = 2014\,Q1,\ 39 = 2022\,Q3$.

Future quarter:	1	2	3	4	5	6	7	8	9	10	11 – 39	Total
Cross-validation	1027	733	465	201	15	5	3	2	1	1	1	2452
Do-validation	970	684	422	166	14	5	3	2	1	1	1	2270
Chain-ladder	948	651	387	148	12	5	3	2	1	1	1	2160

mean structure (Kuang et al., 2009). It predicts a smaller number than the continuous approaches. Under a Poisson approximation with an approximated standard deviation of 48 we get significant differences between the predicted future claims.

8. SIMULATION STUDY

We simulated the two do-validated densities from the application section, shown in Figure 2, assuming the multiplicative structure $f(x, y) = f_1(x)f_2(y)$. These models have been chosen to illustrate realistic situations in claims reserving. Furthermore, for computational reasons, we simulated data by aggregating the occurrence of claims in bin sizes of three days; see the Supplementary Material. We consider four sample sizes corresponding to 0.5, 1.0, 1.5 and 2.0 times the sample size, $n = 58180$, from the application.

For each sample size, we generated 500 samples and have solved the forecasting problem using the methods described in this paper. Since the data are generated in discrete time, the methods were applied using the discrete expressions in the Supplementary Material. The performance of the methods for each simulated data set was evaluated using the discrete approximation of the integrated squared error.

The local linear estimators were calculated using the Epanechnikov kernel with four different bandwidth choices. Firstly the infeasible integrated squared error optimal bandwidth which changes in each simulated sample and secondly the mean of those integrated squared error optimal bandwidths of the 500 simulated samples for every run. These two infeasible choices are compared to the two data-driven bandwidths, weighted cross-validation and weighted do-validation.

Table 3 shows that weighted cross-validation and do-validation perform reasonably well. The results support the asymptotic theory ranking cross-validation as more volatile than do-validation. For the development density, f_2, note that, for larger sample sizes, there is nearly no

Table 3. *Summary of the integrated squared errors multiplied by* 10^5*, along the 500 simulated samples. Four different bandwidths: optimal bandwidth (ISE), averaged optimal out of the 500 samples (MISE), cross-validation (CV), do-validation (DO).*

		f_1				f_2			
n		ISE	MISE	CV	DO	ISE	MISE	CV	DO
29090	Median	0.84	2.45	5.87	6.49	1.40	1.44	1.57	1.50
	Mean	1.50	3.31	18.40	17.65	1.49	1.53	1.66	1.58
	SD	1.72	2.39	33.07	39.64	0.59	0.60	0.68	0.60
58180	Median	0.56	2.29	4.65	4.47	0.84	0.86	0.91	0.87
	Mean	1.12	2.81	11.24	7.21	0.87	0.89	0.95	0.89
	SD	1.30	1.58	17.27	9.09	0.29	0.29	0.34	0.29
87270	Median	0.52	2.42	4.04	3.74	0.62	0.63	0.67	0.65
	Mean	0.99	2.71	7.49	5.29	0.64	0.65	0.69	0.66
	SD	1.14	1.24	11.55	5.68	0.20	0.20	0.22	0.20
116360	Median	0.43	2.35	3.42	3.74	0.49	0.51	0.53	0.53
	Mean	0.89	2.64	5.97	6.15	0.51	0.53	0.55	0.54
	SD	1.06	1.08	8.54	9.00	0.15	0.15	0.17	0.15

difference between the optimal infeasible methods and the two validated bandwidth selectors. In any event, the feasible approaches seem to be doing very well at picking appropriate bandwidths.

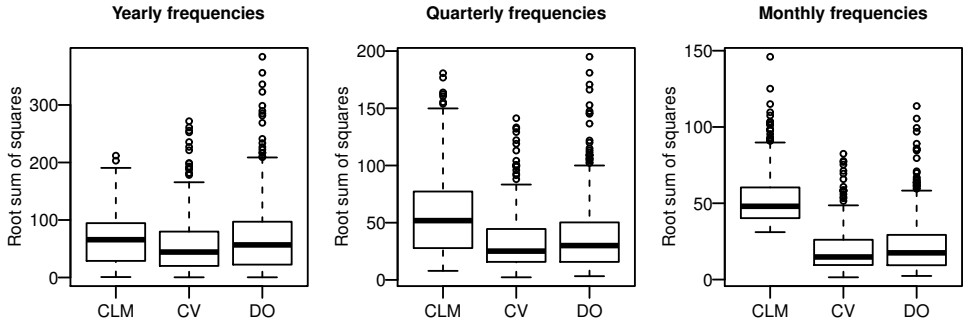
We also simulated the development of the claims according to Table 2. Let R_p be the true reserve for the future period p and $\widehat{R}_p$ its estimator. Then, the error was calculated as $\{\sum(R_p - \widehat{R}_p)^2\}^{1/2}$. Figure 3 shows box plots of the errors in the future count development, obtained from the 500 simulated samples. For comparison, we calculated estimates based on the chain-ladder method, with data aggregated in years, quarters, and months, respectively. Chain-ladder modelling is competitive for yearly numbers, but breaks down for more detailed quarterly, monthly, or daily numbers. It is not included in Table 3.

Our simulations suggest that the do-validation estimate in the application of the previous section, Table 2, is more likely to be nearly correct than the other two estimates. This is supported by the smaller standard deviation of do-validation compared to cross-validation for the sample-size $n = 58180$, Table 3, and by the better performance of the cross-validated and do-validated density approaches in quarterly aggregated data, Figure 3.

9. Concluding Remarks

This paper produces a simpler alternative to the in-sample forecasting approach of Mammen et al. (2015) and Lee et al. (2015). This is done by reversing the time, and it works because all failures are observed until some calendar time. Obviously the simple multiplicative structure of the model could be questioned, see England & Verrall (2002) for some actuarial discussion on the short-comings of the multiplicative chain-ladder model. One possible generalisation of our model would be to let the development density depend on calendar time. Another generalisation would be to include covariates, as has been done e.g. by Wells (1994) for counting process intensities. An example would be to incorporate claim severities. This could be done by extending the counting process set-up of this paper to the marked point processes approach (Norberg, 1993). This could also help to generalise the recent double chain-ladder technique of Verrall et al. (2010), Martínez-Miranda et al. (2011) and Martínez-Miranda et al. (2012) to continuous time. In this paper we developed detailed asymptotic theory for the estimation of the density $f(x, y)$.

Fig. 3. Prediction errors of simulated monthly (right panel), quarterly (middle panel) and yearly (left panel) data along the 500 simulated samples. Sample size is $n = 58180$. Three different methods: Chain-ladder method (CLM), local linear density estimator with cross-validation (CV) and do-validation (DO) bandwidth.



Discussions of plug-in estimators of integrals of the density over triangles and/or diagonals need further theory.

SUPPLEMENTARY MATERIAL

Supplementary material available at Biometrika online includes proofs of Proposition 1 and Theorem 1 and a discussion of discrete approximations of our model.

ACKNOWLEDGEMENTS

We gratefully acknowledge support by Deutsche Forschungsgemeinschaft (RTG 1953), by a subsidy granted to National Research University Higher School of Economics, Moscow, by the Government of the Russian Federation (Global Competitiveness Program), by an IEF grant of the European Community, and by a grant of the Spanish Ministry of Economy and Competitiveness (European Regional Development Fund).

REFERENCES

AALEN, O. O. (1978). Non-parametric inference for a family of counting processes. *Ann. Stat.* **6**, 701–726.

ADDONA, V., ATHERTON, J. & WOLFSON, D. B. (2012). Testing the assumption of independence of truncation time and failure time. *Int. J. Biostat.* **8**.

ALAI, D., MERZ, M. & WÜTHRICH, M. V. (2010). Prediction uncertainty in the Bornhuetter–Ferguson claims reserving method: revisited. *Ann. Actuar. Sci.* **5**, 7–17.

ANDERSEN, P., BORGAN, O., GILL, R. & KEIDING, N. (1993). *Statistical Models Based on Counting Processes*. New York: Springer.

ANTONIO, K. & PLAT, R. (2014). Micro-level stochastic loss reserving for general insurance. *Scand. Actuar. J* **2014**, 649–669.

ARJAS, E. (1989). The claims reserving problem in non-life insurance: Some structural ideas. *Astin Bulletin* **19**, 139–152.

BORNHUETTER, R. L. & FERGUSON, R. E. (1972). The actuary and IBNR. *Casualty Actuarial Society Proceedings* **LIX**, 181–195.

BOWMAN, A. W. (1984). An alternative method of cross-validation for the smoothing of density estimates. *Biometrika* **71**, 353–360.

BROOKMEYER, R. & GAIL, M. G. (1987). Biases in prevalent cohorts. *Biometrics* **43**, 739–749.

ENGLAND, P. D. & VERRALL, R. J. (2002). Stochastic claims reserving in general insurance. *British Actuarial Journal* **8**, 443–544.

FAN, J. & GIJBELS, I. (1996). *Local polynomial modelling and its applications*. London: Chapman and Hall.
GÁMIZ, M. L., JANYS, L., MARTÍNEZ-MIRANDA, M. D. & NIELSEN, J. P. (2013). Bandwidth selection in marker dependent kernel hazard estimation. *Comput. Stat. Data An.* **68**, 155–169.
GÁMIZ, M. L., MAMMEN, E., MARTÍNEZ-MIRANDA, M. D. & NIELSEN, J. P. (2016). Double one-sided cross-validation of local linear hazards. *J. Roy. Statist. Soc. Ser. B* **78**, 1–26.
HALL, P. (1983). Large sample optimality of least squares cross-validation in density estimation. *Ann. Stat.* **11**, 1156–1174.
HART, J. & YI, S. (1998). One-sided cross-validation. *J. Am. Stat. Assoc.* **93**, 620–631.
HEIDENREICH, N. B., SCHINDLER, A. & SPERLICH, S. (2013). Bandwidth selection for kernel density estimation: a review of fully automatic selectors. *AStA Adv. Statist. Anal.* **97**, 403–433.
HIRUKAWA, M. & SAKUDO, M. (2014). Nonnegative bias reduction methods for density estimation using asymmetric kernels. *Comput. Stat. Data An.* **75**, 112–123.
JACOD, J. (1979). *Calcul stochastique et problemes de martingales*. Berlin: Springer.
JACOD, J. & SHIRYAEV, A. N. (1987). *Limit Theorems for Stochastic Processes*. Berlin: Springer.
KUANG, D., NIELSEN, B. & NIELSEN, J. P. (2008). Forecasting with the age-period-cohort model and the extended chain-ladder model. *Biometrika* **95**, 987–991.
KUANG, D., NIELSEN, B. & NIELSEN, J. P. (2009). Chain-ladder as maximum likelihood revisited. *Ann. Actuar. Sci* **4**, 105–121.
LAGAKOS, S. W., BARRAJ, L. M. & DE GRUTTOLA, V. (1988). Nonparametric analysis of truncated survival data, with application to aids. *Biometrika* **75**, 515–523.
LEE, Y. K., MAMMEN, E., NIELSEN, J. P. & PARK, B. (2015). Asymptotics for in-sample density forecasting. *Ann. Stat.* **43**, 620–651.
MACK, T. (2008). The prediction error of Bornhuetter–Ferguson. *Astin Bull.* **38**, 87–103.
MAMMEN, E., MARTÍNEZ-MIRANDA, M. D. & NIELSEN, J. P. (2015). In-sample forecasting applied to reserving and mesothelioma. *Insurance Math. Econom.* **61**, 76–86.
MAMMEN, E., MARTÍNEZ-MIRANDA, M. D., NIELSEN, J. P. & SPERLICH, S. (2011). Do-validation for kernel density estimation. *J. Am. Stat. Assoc.* **106**, 651–660.
MANDEL, M. & BETENSKY, R. A. (2007). Testing goodness of fit of a uniform truncation model. *Biometrics* **63**, 405–412.
MARTÍNEZ-MIRANDA, M. D., NIELSEN, B., NIELSEN, J. P. & VERRALL, R. (2011). Cash flow simulation for a model of outstanding liabilities based on claim amounts and claim numbers. *Astin Bull.* **41**, 107–129.
MARTÍNEZ-MIRANDA, M. D., NIELSEN, J. P. & SPERLICH, S. (2009). One sided cross-validation for density estimation with an application to operational risk. In *Operational Risk Towards Basel III: Best Practices and Issues in Modelling. Management and Regulation*, G. N. von Gregoriou, ed. New Jersey: John Wiley and Sons, pp. 177–195.
MARTÍNEZ-MIRANDA, M. D., NIELSEN, J. P., SPERLICH, S. & VERRALL, R. (2013). Continuous chain ladder: Reformulating and generalising a classical insurance problem. *Expert. Syst. Appl.* **40**, 5588–5603.
MARTÍNEZ-MIRANDA, M. D., NIELSEN, J. P. & VERRALL, R. (2012). Double chain ladder. *Astin Bull.* **42**, 59–76.
NIELSEN, J. P. (1998). Marker dependent kernel hazard estimation from local linear estimation. *Scand. Actuar. J.* **1998**, 113–124.
NIELSEN, J. P., TANGGAARD, C. & JONES, M. C. (2009). Local linear density estimation for filtered survival data. *Statistics* **43**, 176–186.
NORBERG, R. (1993). Prediction of outstanding liabilities in non-life insurance. *Astin Bull.* **23**, 95–115.
R DEVELOPMENT CORE TEAM (2016). *R: A Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing. ISBN 3-900051-07-0, http://www.R-project.org.
RUDEMO, M. (1982). Empirical choice of histograms and kernel density estimators. *Scand. J. Stat.* **9**, 65–78.
SAVCHUK, O. Y., HART, J. D. & SHEATER, S. J. (2010). Indirect cross validation for density estimation. *J. Am. Stat. Assoc.* **105**, 415–423.
TSAI, W.-Y. (1990). Testing the assumption of independence of truncation time and failure time. *Biometrika* **77**, 169–177.
VERRALL, R., NIELSEN, J. P. & JESSEN, A. (2010). Including count data in claims reserving. *Astin Bull.* **40**, 871–887.
WANG, M.-C. (1989). A semiparametric model for randomly truncated data. *J. Am. Stat. Assoc.* **84**, 742–748.
WARE, J. H. & DEMETS, D. L. (1976). Reanalysis of some baboon descent data. *Biometrics* **32**, 459–463.
WELLS, M. T. (1994). Nonparametric kernel estimation in counting processes with explanatory variables. *Biometrika* **81**, 795–801.